Bryn Mawr Classical Review

Bryn Mawr Classical Review 2001.09.23

TLG, *Thesaurus Linguae Graecae, CD-ROM Disk E*. Distributed by the Thesaurus Linguae Graecae for the Regents of the University of California, February 2000. ISBN 0-9675843-0-2.

Reviewed by P.J. Heslin, University of Durham (p.j.heslin@dur.ac.uk)
Word count: 7210 words

Last year the *Thesaurus Linguae Graecae* (TLG) published on CD-ROM the latest update to its ever-improving database of Greek texts. The chief feature of this update is the addition of thousands of new texts, primarily post-Classical; this has increased the size of the database from 458 to 650 megabytes in total. A full list of the authors and works making their first appearance in the database and those authors furnished with new or revised texts is available from the TLG web site.[1] The TLG has sent invitations to current subscribers to exchange their old version, disk D, for disk E at no additional cost, but a reader of classical interests might wonder how much benefit the upgrade will afford someone not primarily concerned with texts from the Christian period.

In fact, disk E brings very many important improvements to the quality and number of texts of Classical interest that the TLG contains. An example might illustrate this better than a long list of new texts: let us imagine that a reader of Homer comes across the word κλυτ□πωλος, an epithet of Hades, and is curious as to what Hades was thought to have to do with horses. The old D version of the TLG already provided a great wealth of comparative material: 22 instances of the word in 9 authors. We find it in the fragments of Hesiod and of Pindar (who uses it of Poseidon, a likelier horse-fancier). Then there are the grammarians and the commentators on Homer and Pindar: Apollonius Sophista, works going under the names of Tryphon and Aelius Herodianus, the Homeric scholia, Eustathius, and Hesychius. This is a much fuller picture than can be obtained from a lexicon, but now one can do even better.

When we run the same search on disk E, we get all of the citations reported by disk D plus 13 more for a total of 35. The 5 new authors include one that adds nothing new: Apollonius Sophista's quotation from the Homeric glossary of Apion now appears under Apion's name as well as Apollonius'. The other four new authors are more interesting. Now the result includes Tryphiodorus, a late author, but the only non-Homeric citation offered by the lexicon of Liddell-Scott-Jones: his connection of the epithet with the episode of the Trojan horse is the most interesting semantically of all the citations. Then there is a cogent entry in the *Etymologicum Magnum* which anticipates the comments of modern scholars (eg. Kirk ad *Il.* 5.652-4) in referring to Hades' chariot and the rape of Persephone. That explanation is also offered in the Homeric D-scholia, but they were not among the scholia found on the D disk; the D-scholia are an example of a new text added to an old author in the E disk.[2] Finally, there are two more new texts that are quite different in kind, but which both apply the epithet to celestial phenomena. An hexametrical work on weather (Maximus Astrol.) uses it three times as an epithet of the moon, while the magical papyri preserve an example of its use in a prayer to the sun. This small example may serve to show that, quite apart from the multitude of church

fathers, there is much else that is new and useful in the latest TLG.

Particularly valuable for unexpected juxtaposition is the addition of texts such as the magical papyri (ed. Preisendanz and Henrichs). Many minor but interesting authors such as Tryphiodorus and Musaeus appear for the first time, and so do several whole minor genres, such as the *periplous* narratives, alchemical writing and paradoxography. Many scientific and medical writers have been added, including such major figures as Archimedes, Claudius Ptolemaeus and Artemidorus. There are also very many new collections of fragments and editions of the fragments of individual authors, and important grammatical works, such as the various collections of etymologies. The new disk takes advantage of important recent editions, such as Stephens' and Winkler's fragments of the Ancient Greek novels, Irigoin's Bude/ of Bacchylides and Obbink's Philodemus *On Piety*, all from the mid-1990's; it includes texts published as late as 1998.

Euripides and Bacchylides are examples of authors who have long been a part of the corpus but have been provided in Disk E with new texts. The change from Murray's edition to Diggle's is not the only improvement in the text of Euripides: a comparison of the old text-file with the new shows how much the TLG has learned about the process of digitization. The old Euripides is marked up in a very fussy manner, with more detail recorded about the placement of the type on the page than anyone could reasonably desire; the result is that when the file is viewed directly it is hard to find the text amid the mark-up. The new Euripides has more limited formatting information, and is much easier for humans, and presumably for computers, to parse. Before further discussion of such technical details, however, it may be helpful to take an overview of how the TLG works.

The Implementation of the TLG

No retrieval software is provided with the TLG CD-ROM, but rather the user is expected to obtain this separately; the license states that details of how the database works will be provided on request to enable the writing of such software. I have never requested nor received this information, so the reader should bear in mind that the present review is based on my own subjective impression of how the TLG was designed to work, and that at any point I may be mistaken in my surmises and guesswork. The TLG web site has long had basic information on the Greek encoding and page mark-up scheme (Beta code), and this has lately been supplemented by a much more detailed and helpful account; a series of technical papers of very high quality have also appeared on the same site quite recently. This seems to have been in part a side-effect of the development for the first time by the TLG of its own retrieval software for general use: this is the new online interface to the texts.[3] The proper way for the TLG to ensure that reviews such as this one are as accurate and fair as possible is to eliminate guesswork by disseminating publicly the complete documentation on its use rather than by private arrangement. I hope that the publication of the new documentation on the TLG web site signals a trend in this direction.

Another caveat to bear in mind is that, since there are many different pieces of software for accessing the TLG, search results obtained may vary slightly from person to person. Despite this disadvantage, the policy of distributing data only, and leaving users to create or purchase their own software for accessing it, is precisely the correct one for a serious scholarly project. Contrast many other databases, such as the *Bibliotheca Teubneriana Latina*, which can only be accessed via vendor-supplied software. This means that it cannot be peer-reviewed; that you are restricted to using a certain type of computer and operating system to access the database; that you can only ask it the type of questions that have been envisioned by the designers and allowed by their software; and that you cannot easily and independently verify the correctness

of the responses given by the software. In a recent BMCR review, Anne Mahoney made the point more generally: "Electronic texts are most valuable when they can be used by other scholars, not just their creators, perhaps for purposes the creators did not anticipate ...".[4] It is the open architecture of the TLG that makes it possible for scholars to use it creatively, and this is one of its most important features.

In order to make some technical points about the changes in Disk E, it will be necessary to explain in some detail how the data are structured. This may be a useful exercise in its own right since traditionally there does not seem to have been a great deal of publicly available information on many of these details. With the exception of reading the recently published TLG technical notes, I simply proceeded by examining the data on the CD-ROM. All of the results reported here have been generated by software of my own devising.[5] As a consequence, this review is written from a rather low-level perspective, but I would seek to excuse the tedium of the technical material that follows on the grounds that some users will, I think, feel better able to use the TLG effectively if they know more about how it works. Others may not care; they should skip the next 3 sections of this review, and continue with the section entitled "Has anyone around here seen the TLG disk?".

File Formats

If one were to start the TLG project over again today, the polytonic Greek encoding to use might be the Unicode standard, and the recording of document structures might be done in XML; but these technologies have only recently emerged as standards. As a result the data in the TLG are in an idiosyncratic format, which was designed for maximum compactness in the days of very limited computer memory. It is a tribute to the intelligence of the basic design that it has aged so well: search performance has continued to increase along with the development of faster computers, without being hindered by the original assumptions behind the data format.

The TLG comes on one standard CD-ROM, containing text files, index files and support files. The structure of the text and index files is shared with the publications of the Packard Humanities Institute (PHI). Each text file contains all of the works of a single author (or collective group), and each is named according to the four-digit author number given in the TLG canon:[6] thus the texts of Homer, designated author 0012, are found in a file called TLG0012.TXT. The text files contain the author's works themselves and coded markers for citation information (line numbers, etc.) that are integrated with the text in a very compact way.[7] Each text file (.TXT) has a companion index (.IDT) file that indicates among other things the name of the author, the titles of his or her works, and what terms should be used to label the citation information (book and line for Homer, Stephanus page, section, and line for Plato); for Homer this file would be called TLG0012.IDT.[8] The support files include among others an index of all of the authors contained on the disk (AUTHTAB.DIR); digital versions of the canon of texts, giving full bibliographical information; a set of files categorising authors and works by date, genre, location, gender and epithet;[9] and the files containing the word list, discussed below.

BMCR readers should be familiar with Beta code, the TLG transliteration scheme, since that is what is used by this journal. Its most annoying feature is the exclusive use of Latin upper-case letters, with Greek upper-case indicated by an asterisk before the letter. The Latin lower-case letters made free for use by this rule are used for nothing at all except for the additional letters in the Coptic alphabet of demotic origin (these texts are on the PHI disk). If that is why squinting BMCR readers are forced to abandon centuries of progress, reverting to the single-case poverty of the ancient Greeks themselves, then it is a strange case of the Coptic tail

wagging the Greek dog. Another annoyance is that accents on upper-case Greek letters are encoded before the letter, where they appear typographically, rather than after the letter, as with lower-case letters; this makes case-insensitive searching and alphabetization even more of a challenge. For all of its faults, the Beta encoding has been around for a long time and so has established itself as a standard.

The Word List

A feature of major importance in the TLG is the word list. It is often not practicable to search the entire database, particularly on a slow computer, and so a means is provided to speed up the process: a list of all of the words (and indeed of all runs of several characters) found in the corpus along with information on which works in the TLG contain each word.[10] An important limitation on this provision is that there is no information on exactly where in each work the word is to be found; this means that one still has to take the time to find it within each given work. The advantage is that one does not have to waste time searching in those works that do not contain the word; thus the advantage is greatest in searching for rare words and it is nil in searching for a word that occurs in every work. As searches grow faster, the urgency of using the word list diminishes, and with a reasonably fast desktop computer it is possible today to search the entire TLG corpus, even at the expanded size of Disk E, in about three minutes without using the word list, which is tolerable.[11] Nevertheless, many people who have always been accustomed to using the word index as a convenience will continue to do so. If so, there are some issues they should consider.

An earlier generation of computers could not hold the entire word list in memory, so an index to the list is also provided (in a file named TLGWLINX.INX). This index uses the first two letters of each word to point to the appropriate section of the word list, which is ordered alphabetically. Depending on whether or not your retrieval software uses this supplementary index, your search of the word list may depend upon the first two letters of your query. Thus a search for the word □γαθ□ς might return the form κ□γαθ□ς if your software searches the word list directly for matches, but it will not find it if it uses the index to the word list, since that word begins with a different letter.

Difficulties grow when you consider the many problems of morphology. A general search in Disk E for words beginning with αγαθ yields 489 forms, which is a lot to weed through if you are only interested in inflected and not in derivative forms of □γαθ□ς. Even then, you may have to remember to look separately for variants like κ□γαθ□ς. For verbs the problem escalates further, and for irregular verbs it would be quite a job to ensure that one has thought of every possible inflected form the word may take. There is no easy solution to that problem, but one possible approach, the development of a lemmatized index, is described below.

The compressed format of the word list data has permitted a great deal of information to be packed onto a single CD-ROM, but it also makes it hard to find errors. There is one such error in the word list on Disk E, where the count of the word κα□ has gone awry. If you look at the word list, it says that κα□ appears a total of 4,177,357 times, but if you add up the per-work totals for all of the texts in the TLG the sum is 4,111,821. The difference is 65536, which is a very round number in the binary arithmetic that computers use: 2 to the power of 16. It thus looks very much like a overflow error, typical of computers (the year 2000 bug was just such an error, occurring when an insufficient number of digits has been made available for recording a number). Robin Smith has pointed out to me the likely site of this overflow error: it is in the *Iliad* commentary of Eustathius. According to Disk E, the number of times that work repeats the word κα□ is 1108, which is much, much too low. If we compare the same text in Disk D, κα□ is said to appear 65535 times, which is more plausible; but it is just one less than our

suspiciously round number.

It appears that what happened in Disk D is that the word list simply stopped counting when the number of times a word appeared in a work hit the maximum allowed in 16 binary digits, or 65535; fortunately this only happened in one very long work. In Disk D, the overall wordlist and the Eustathius data thus agreed with each other but both were wrong. In Disk E the overall word list appears to be correct, but the Eustathius data, instead of stopping at the maximum, went back to zero and began counting again from there, until it reached 1108. I would guess, then, that the correct number of times Eustathius used the word κα□ in his *Iliad* commentary is 1108 + 65536, or 66644. It may be that when searching for the word κα□, some retrieval software will be confused by the discrepancy between the total in the word list and the total of all of the per-work data, but otherwise it is not a serious bug. Still, it would be good if the TLG could find a way to make room for more than 16 binary digits in the per-work incidence data without breaking compatibility with older retrieval software. If that proves impossible, an alternative would be to split the Eustathius commentary into two halves.

Another odd bug I found applies to one particular work of Maximus Confessor, his "Expositio in Psalmum LIX" (author 2892, work 110). It seems to me that the word-list data claim that every word in that work appears exactly twice as often as it actually does. The only exceptions to this are hyphenated words: they appear only once in their correct and integral form; however, the separate pieces of each hyphenated word also appear in the word list as distinct entries. I have not done a comprehensive test of all of the works in the corpus to see whether this problem recurs elsewhere. Reporting too many instances of a word is not as serious a problem as reporting too few; but it may give rise to some software errors. The strange behavior of hyphenated words in this work suggests that the origin of this bug might be connected with the reworking for Disk E of the policy by which the word list treats hyphenated words, as described below.

Once you have found the word you want in the word list and have determined the works in which it may be found and how often it is found there, a more general difficulty arises. When you look for that word in the texts, it may appear there in hyphenated form, elided, with an extra accent thrown back from a following enclitic, etc. Since the TLG tells us only that a word appears x number of times in a given work and does not tell us exactly where, it is important to know precisely what constitutes a word for the purposes of the word list and how it is transformed into its canonical form, so that we can search for exactly those same words. Fortunately the TLG has recently published on its web site clear and comprehensive documentation of what constitutes a word; this definition has become much more complex with Disk E.[12] For example, it says that for the purposes of the word list, "Accents are regularized: grave accents become acute, only the final accent in a word is retained, and words are converted to lower case." This seems reasonable enough, but it sometimes has strange consequences; for example, if a word has two accents because an editor has supplied an alternate reading, it may happen that the brackets are removed, both vowels appear, but the second has its accent removed.[13]

The appearance of oddly-formed words in the word list is not really a major problem, especially since fragmentary works yield all sorts of odd quasi-word artifacts; but the business of finding these in the texts can be hard. Even with a written specification of what constitutes a word, there are many issues that emerge from the transition from Disk D to Disk E, especially with regard to hyphenated and fragmentary words; some of these changes are deliberate, others may not be. For example, if we do a search for the word □π□λλωνος in Callimachus, we find 11 examples in Disk D but only 10 in Disk E even though both disks use the same electronic text (slightly rearranged).[14] The missing word in Disk E has become now .□π□λλωνος, where

the exclamation point denotes the place where a word-fragment has broken off. At line 8 of Iambus 1 (fr. 191), Pfeiffer did place a dot indicating the vestige of a stroke just before the name of the god; since no space intervenes, this, according to the new technical definition, constitutes a word-fragment, and is so considered by Disk E. On the other hand, Pfeiffer capitalized the god's name, so he clearly meant that it was a complete word, and that is what Disk D considered it, perhaps inconsistently, but correctly.

The rules according to which partial words are recognized have been extensively reworked and made much more sophisticated, and retrieval software will have to be changed to match these new definitions. For the most part these changes are sensible, but they cannot be regarded as final, as there are still problems in the handling of hyphenated words. One of the changes is the introduction of rules whereby part of a word at line-end, ending in a hyphen and followed by a lacuna on the next line, is considered quite rightly to be a word-fragment. The problem is that there is a large number of words where a hyphen is followed on the next line by a quotation mark, andthen the rest of the word. This arises from the old typographical convention whereby multi-line quotations were displayed with an opening quotation mark at the beginning of each new line of the quotation. This style of quotation (which has had a revival of sorts in the conventions of e-mails) is quite frequent in some of the older printed texts used as exemplars by the TLG, especially in ecclesiastical works (see for example the commentaries of St. Cyril of Alexandria). In Disk E it seems that every word that is hyphenated across one of these line-beginning quotation marks appears in the word list not as a single word, but as two separate pieces. This is clearly not what the TLG specification intends, as hyphenated words elsewhere are always joined together. It appears that a line beginning with a quotation mark in front of the resumption of a hyphenated word is mistakenly taken to indicate a lacuna. The D Disk handled these types of quotations correctly. A good deal of work has gone into the more accurate recognition of partial words from fragmentary texts, and this goal is worthwhile, but it should not interfere with the handling of normal hyphenation in integral texts.

These subtle changes in how exactly a word is defined should not make a difference, but they do. The problem is that the TLG decides what a word is, and then leaves it to retrieval software to search for those words using precisely the same criteria that the TLG itself used. Yet we have seen that from version to version subtle changes may be introduced in the definition of a word; tracking these changes while maintaining backward compatibility with older versions of the TLG disk is a hard task for retrieval software. If the TLG were to publish a word list that included not only per-work totals, but also the locations of the words within the texts, then the problem would be eliminated for those using up-to-date retrieval software. The TLG could fiddle with its definition of a word at will since the retrieval software would not have to search for the words nor care how they are defined, but merely jump to the right location. A further advantage would be that searches would be instantaneous. Apparently, this is the strategy pursued by a piece of retrieval software created at the Scuola Normale Superiore,[15] for the purpose of speeding up searches. This sort of supplementary index could be distributed on one or more extra CD-ROMs, and might exclude the locations of the most common words. This is information that the TLG has to generate in any case, and it seems that it is currently using this full indexing data, as would be expected, in the mechanism for doing searches in the new online TLG. If, as seems likely, the TLG will in the future have to distribute multiple CD-ROMS, why not provide this data too? Alternatively, the TLG might suggest the terms under which other makers of retrieval software might generate and distribute such indices of their own.

TLG Formatting Codes

Apart from the texts themselves and the encoded citation information (e.g. book 1, line 611)

embedded alongside, the .TXT files contain a great deal of page-formatting information that is encoded by using analphabetic characters. Some of this information pertains to content (e.g. "this is a stage direction") but most of it is visually oriented (e.g. "this is printed in small caps, in the margin"). The intent of this mark-up is thus a bit muddy, but usually the goal seems to be to represent the appearance of the printed page as closely as possible, even where, as in the case of hyphenation, it makes the text more awkward to manipulate electronically. This is in stark contrast to the usual practice in most modern text encoding schemes, where the intent is to capture the abstract structure of the document rather than the accidents of its appearance in a printed text. Presumably the benefits of this approach are that it is easier visually to confirm the accuracy of the transcription against the exemplar and that it does not require the transcriber to master the sometimes complex and varying typographical conventions of textual criticism and papyrology. The disadvantage is that it shifts onto the user the burden of coping with hyphenation, with shifting accents on capitalized words, and with making sense of the text amid a sea of obscure mark-up symbols.

The analphabetic characters that specify formatting information are themselves part of a hybrid scheme: many types of formatting are indicated by a beginning tag and an ending tag, analogous to HTML, whereas the font information and switches from Greek to Latin text operate quite differently. Changing fonts are indicated by a command that says "switch fonts here", but which does not say for how long that particular font is to have effect. The problem is that you can start a special effect and forget to specify where it ends; thus formatting that is meant to be local can easily have non-local effects.[16] Contrast a tag-based formatting scheme, such as that used for other types of formatting in the TLG, or such as HTML, where a tag that begins italics *must* in theory be matched by one that ends it; cases where a matching tag is missing can be identified easily by an HTML-validating program.

Consider this bit of the TLG, from the fragments of Chaeremon:

> @&10dramatis personae%10 [1$10*P*A*N

According to the Beta code document on the TLG web site, this means: "Indent (@), then switch to a small Roman font (&10), then print the words 'dramatis personae', then a dicolon (%10), then a space, then the opening of a parenthesis ([1), then a small Greek font ($10), then the word 'PAN' in upper-case, as indicated by the asterisks." The cast of characters continues, with further modifications of the current font, but at the end the instruction to convert to a small Greek font is never matched by an instruction to begin a normal-sized Greek font. Thus all of the subsequent text is wrongly formatted at a small size, until another bit of text happens to come along with a command to switch to a normal size at its end.

The present scheme was presumably adopted for motives of economy of space: indicating italics the way the TLG does, &3like so&, takes three extra characters; the way HTML does it, *like so*, takes seven. The disadvantages are many: it is obscure, whereas tag-based schemes are more obvious and self-documenting; it is difficult to validate; and it is difficult to convert to a more generic and completely tag-based mark-up that might be displayed by a browser. The awkwardness of the interaction between the font information, which is state-based, and the rest of the formatting, which is tag-based, may be illustrated by the very beginning of Erbse's edition of the Homeric scholia:

> @{2& D %5 ex. %5 ex.$}2

This is the marginal comment "D | ex. | ex." that serves to identify each of the sources of the three sections from which the first scholion is compounded. The mark-up means: "Indent (@),

then begin marginal note ({2), then switch to a normal Roman font (&), then 'D', then a vertical bar (%5), then 'ex.', then another bar, then another 'ex.', then switch back to a normal Greek font ($), then end marginal note (}2)." The indentation at the start refers to the beginning of the main text, not the marginal note, and this is logical if we assume that the brackets define the scope of the marginal note and that whatever lies outside of them pertains to the main text. The last symbol that lies within the marginal note, however, is a change of font state that is to take effect outside the note: the character $ at the end of the note indicates a switch from Latin back to the Greek text in which the scholion itself is written. Thus the font switching mechanism is completely independent of the rest of the formatting, and the division between what is inside and outside the note is less than clear-cut. In consequence, if one tries to write a system that will display this Beta code by transforming it into a more standard type of mark-up, it is very difficult to cope with local areas of Beta code that are part of a piece of text printed quite separately from the main block, but which nevertheless determine the font style of the main block.

In these days of computers with large memories and storage capacities, the eventual conversion of the TLG texts to a less idiosyncratic format would greatly facilitate the display of all of the rich typographical detail captured in the database. All too often, retrieval software simply throws away a large fraction of this formatting information as too difficult to grapple with.

Has Anyone around Here seen the TLG Disk?

The most unsatisfactory aspect of the TLG is its underutilization by those who could benefit from it. When I was an undergraduate, using the TLG involved approaching a member of staff who knew how to use the Ibycus microcomputer; I never had the courage to ask. As a postgraduate, it involved me in signing out the CD-ROM from the department, finding a free computer in a lab, installing retrieval software and a Greek font, and sitting there for forty minutes or so while the search proceeded; I did it when I absolutely had to. Now I can hoard my department's CD-ROM in my office; a fine solution for me, not so good for my colleagues and our students. Part of the reason for the current inconvenience of accessing the TLG is one particular stipulation of the Individual and Institutional License, which is otherwise perfectly reasonable. It is absolutely forbidden to copy the data from the CD-ROM onto the storage of even one computer, even for the purpose of putting the CD-ROM away for safekeeping. Even commercial software firms recognize that while a CD-ROM is well suited to the task of distributing large amounts of data cheaply, it is a very inconvenient medium for frequent access. Your word-processor does not insist that you load the program from the CD-ROM you bought it on every time you start it up. This regime, whereby the CD-ROM itself, rather than the valuable data on it, is the focus of the license, can lead to overuse, damage and loss of the disk. It also means that anyone who has the notion of using the TLG has to overcome the nuisance factor inherent in finding both a free computer and the disk itself. This strange situation is reflected in the policy on damaged and lost CD-ROMs as documented on the TLG web site.[17] We are told that, "[w]hile CD ROMs are relatively rugged data storage devices, they can be damaged by careless handling (e.g., dropping, improper insertion into a compact disk unit, or scratching of the surface)." CD-ROMs are said here to be both rugged and easily damaged; the contradiction arises because the TLG is requiring users to subject their CD-ROM to the sort of constant handling by many people that is not good for it. Even more curious is the assertion that follows: "Like an expensive printed volume which--if lost or stolen--will not be replaced by a publisher free of charge, a lost or stolen TLG CD ROM cannot be replaced at the TLG's expense." That, however, is precisely the point: an electronic database is not at all like an expensive printed volume in terms of its tangible form. Large printed volumes are very expensive to produce, print, warehouse, and ship. Databases are very expensive to produce, but cost next to nothing to reproduce, store and ship. Students cannot use their department's copy

of the *Thesaurus Linguae Latinae* as a frisbee; and you are unlikely to ruin your entire run of the *TLL* by inadvertently using it as a coaster for your coffee mug, or let it slip unnoticed behind a filing cabinet. Databases do have the advantage, however, that a copy can easily be made for everyday use while the original is kept safe, provided that the license permits.

There are technical reasons, too, why insisting on keeping the TLG data on the CD-ROM is not sensible. Currently available CD-ROM drives are running at the maximum speed they can spin without the centrifugal force degrading the physical structure of the plastic disk. Since the data on a CD-ROM is read sequentially, this imposes a ceiling on how fast the database can be accessed. Hard disk drives, on the other hand, continue to improve their access speeds. Once upon a time, CD-ROM disks could hold far more data than the hard drive on the average desktop computer, but the situation is very much the reverse today. The sensible thing would be for the TLG to follow industry standard practice, and permit the individual and institutional licensee to copy the data onto one and only one computer (or two, or three, according to the license paid for), provided that the CD-ROM was not simultaneously used, and provided that access was not provided to other users over a network. Users would have faster and more convenient access, and the TLG would have less trouble with lost or damaged disks.

The License

There are three licensing options for the TLG: the Individual License for US$ 300 for five years, Institutional License for US$ 850 for five years and a Site license. The Site license is the only one that permits downloading the data to a hard disk and network access (it is odd that these two items are consistently linked in the TLG licensing documents, as though one implied the other); it also includes access to the online version of the TLG. The pricing scale for this license is not stated on the web site, perhaps in deference to the sensibilities of less generously funded institutions (is the message that, if you have to ask, you can't afford it?). As for licenses for private individuals, US$ 300 would be expensive enough as a once-off payment, but as a five-year rental at the end of which period you have to pay up again or stop using the database it will be quite out of the question for most people. The idea of a five-year rental is fine for institutions that can project budgets for essential items like the TLG, but as an individual convenience it will appeal to very few at this price and on these terms. If the TLG is serious about putting its database within the reach of individuals, it will need to rethink this policy, which seems to hark back to the days when a computer sufficiently powerful to use the TLG efficiently was rarely the property of, say, a graduate student. One simple possibility is to sell the CD-ROMs to private individuals outright. Such a license would not include automatic upgrades, but it also would not expire after five years; updates might or might not have to be purchased, perhaps at a discount over new. This is hardly a radical idea; it is how the software industry works; it is more or less how book publishing has always worked. The TLG has to work within the agreements it has with the copyright holders of the texts in the corpus, but even so, I think they could do better at coming up with ways to expand access and generate revenue at the same time. As for institutions, it might be worth considering the provision of some intermediate ground between the extremely restrictive Institutional License and the much more permissive Site License. For the many academic departments with a small computer room, a per-seat license allowing only local network access might be attractive.

Future Development

With the release of Disk E, the TLG corpus now contains 650 megabytes of data, which is very close to the maximum amount of uncompressed data a typical CD-ROM can hold, so the question of how future versions will be released presents itself. In fact, the TLG has continued to gain additions even after the release of Disk E, and access to these new texts is available to

those with a license to use the online version. This is a nice bonus for institutions that can afford the Site License, but I certainly hope that it does not mark a shift away from distributing an up-to-date TLG on CD-ROM. Increases in the power of personal computers have meant that performing a search of the TLG has gone from being an afternoon's occupation to being even quicker than looking up a word in a lexicon. It would be a shame to lose that flexibility in exchange for a single interface and a single access point subject to the potential slowness of network connections and overloaded servers. Assuming that the TLG does not want to shift away from the CD-ROM to a higher-density medium and thereby break compatibility with older access machinery like the Ibycus microcomputer and most current desktop computers, the natural route to take would be to distribute the TLG on multiple CD-ROMs. This is the route taken very successfully by the so-called platform-independent Perseus, version 2.0. It comes packaged in four CD-ROMs that can be loaded together onto a single hard disk drive, or, for older computers without that much free storage space, each disk can be accessed independently. The same strategy would work for the TLG, too. The corpus could be divided into, say, a pre-AD 600 disk and a post-600 disk; each disk could be used separately, even on the oldest machines. In order to use the corpus as a whole, you would copy the contents of both CD-ROMs onto a hard disk, and then add the support files for the entire corpus from a third CD-ROM. If the TLG were distributed on multiple CD-ROMs, another convenience that would be made possible is the provision on one or several extra CD-ROMS of the exact data on the locations within the texts of the words in the word list.

The greatest improvement that could be made to future editions of the TLG, however, is the provision of a lemmatized index to the word list. This would be a file containing a list of dictionary forms of Greek words and with each word pointers to the locations of its various inflected forms in the word list. The most obvious benefit of this would be that you could search for all instances of φ□ρω in the author of your choice without having to think out all of the possible forms it might take. Another benefit would be that analysis of word-frequency statistics could be done much more easily. For example, my opening example of a reader searching for the word κλυτ□πωλος was not plucked out of the air. I noticed that one of the major changes in Disk E was that the magical papyri kept cropping up in unexpected places, so I performed a search for all words in the word list that occur in Homer and the magical papyri, and in no more than 15 other authors. Most of the results were not really what I was looking for, since they were odd inflected forms of much more common words. The isolation of interesting words on the basis of usage patterns would thus be one of the less obvious benefits of a lemmatized word list.[18]

Generating this index in a comprehensive form would not be a trivial or automatic task, but it might not have to be begun from scratch nor implemented all at once. Several access methods for the TLG provide links from each Greek word to the online morphological analysis tool of the Perseus project; what the TLG itself needs is to map that information in the other direction; that is, it needs to provide a way to go from dictionary form to inflected form, rather than vice versa. Perseus has far fewer texts than the TLG, and not all of the words found in all of the TLG texts will be known by the Perseus morphological tool, but even a partial implementation of this sort of index would be useful. If the word you are looking for were not available, then you would simply have to revert to looking for the inflected forms by hand.

In summary, the TLG has long been recognized as one of the most ambitious and successful digitization projects in the world, and the latest version proves that it has not been resting on its laurels. Among the important advances that Disk E has brought are many new texts, better versions of old texts, an improved implementation, and vastly better documentation. The only major criticism that can be made is that such a valuable tool is not as widely or as conveniently used as it could be. This situation could be addressed by providing a lemmatized, partial index

to the word list, by providing the word-location data to accompany the word list, and by rethinking parts of the licensing structure to bring the database closer to the fingertips of its users. The possibility exists of making the TLG as fast and convenient to access as a conventional lexicon or thesaurus, and I am sure that those who have brought this project on such a successful journey thus far will be able to meet the new opportunities for access provided by the continuing proliferation of inexpensive and fast desktop computers.

NB. All WWW sites were accessed on 15 August 2001.

Notes:

1. The web site is at http://www.tlg.uci.edu/, and the list is at http://www.tlg.uci.edu/CDEworks.html.
2. The Homeric scholia are grouped together as one "author", with the various collections of scholia included as separate works. The complicated name given to the D-scholia in the TLG E disk, Scholia in Iliadem (scholia vetera) (= D scholia), may be a bit confusing, since Erbse in fact excluded the D-scholia, which he called *scholia minora*, from his edition of the *scholia vetera*, which is why they did not appear in disk D of the TLG.
3. The old, basic description of Beta code is available at http://www.tlg.uci.edu/~tlg/BetaCode.html; the new, comprehensive document is at http://www.tlg.uci.edu/~tlg/BetaCode.pdf. The TLG technical notes 001 - 004 are at http://www.tlg.uci.edu/~tlg/help/Doc001.html, etc. This excellent new documentation, written by N. Nicholas, was sorely needed and is clear, comprehensive, and illustrative. There is a demonstration version of the online interface with a very limited number of TLG texts available at http://www.tlg.uci.edu/~tlg/demo.html.
4. BMCR 01.07.10 (10 Jul 2001), Hockey, *Electronic Texts in the Humanities*, reviewed by Anne Mahoney.
5. Available from http://www.durham.ac.uk/p.j.heslin/diogenes.
6. *Thesaurus Linguae Graecae Canon of Greek Authors and Works*, Berkowitz and Squitier (Oxford University Press, 1990), with updates from the TLG web site; digital versions are included on the TLG disk.
7. Technically speaking, the text files contain lines of ASCII text as variable-length records separated by citation state-update information encoded as runs of non-ASCII bytes; these records are packed into 8 kilobyte, zero-padded blocks. Full citation information is reiterated at the start of each block, providing a good measure of robustness and the possibility of fairly random access.
8. One improvement that could be made to the design of the .IDT files would be to include abbreviated bibliographical information about each work, so that the user does not have to make a separate search in the Canon to discover that the obscure author his search has turned up was in fact taken from a collection of texts close to hand.
9. Some of the data contain strings with extra spaces appended at random, such as "femina ", which is the sort of thing that can lead to difficult-to-diagnose software errors.
10. The word list is in the file TLGWLIST.INX along with total counts for each word, and the file TLGWCNTS.INX contains the encoded information on how many times each word is found in each work of the TLG. The minimum length of a word-fragment changed in Disk E: for details see http://www.tlg.uci.edu/~tlg/help/Doc001.html.
11. Seaching for κλυτοπ□λ, the example used above, took 3 min, 5 sec on a 733 MHz, 128 MB Pentium III with a 52x CD-ROM drive, running Linux 2.2.17 and using the command-line version of Diogenes (version 0.93); the same search using the word list took less than 15 seconds.
12. See "N. Nicholas, TLG Technical Note 001: Greek Word Definition" at

http://www.tlg.uci.edu/~tlg/help/Doc001.html.
13. For example, see βρυχ□[□]μενος in one version of Tzetzes' Scholia on Aristophanes' *Frogs* (author number 5014, work 023), which appears in the word list as βρυχ□ομενος. The new word-list specification tries to deal with this problem intelligently, but the variety of ways that editors may indicate corrections and alternate readings and the variety of ways these have been encoded by the TLG makes it hard to catch all instances.
14. The text of Callimachus is one of those that has been updated in Disk E: whereas Disk D took all of the poet's fragments from Pfeiffer's edition and considered them as one work, with additional fragments from the *Supplementum Hellenisticum* entered as another work, now the *Aetia*, *Iambi*, etc. from Pfeiffer's edition are considered to be separate works.
15. I have never tried it, but there is some information at http://www.cribecu.sns.it/~snsgreek/ENG/feat.htm.
16. Technically speaking, font information is implemented as a state machine. At any given point, the font formatting is in a particular state: for example, oblique small Greek type. The next formatting code simply sets the next state without reference to what went before: a symbol for normal Latin font would change the current state to normal Latin text; this format would obtain until the state changes again. This is similar to the way that the changes in the citation information are encoded. The arrangement works well for the citation information, which has a limited number of states that are changed in a very regular manner. For the messier problem of text mark-up, it doesn't always work so well.
17. http://www.tlg.uci.edu/Damaged.html.
18. One of the help files for the online TLG notes that the TLG is not lemmatized "[a]s of this writing", which I hope indicates some interest in that direction (http://www.tlg.uci.edu/~tlg/help/HelpLex.html).

Read Latest	Index for 2001	Change Greek Display	Archives	Books Available for Review	BMCR Home

HTML generated at 13:27:31, Friday, 03 April 2009